BAB II
TINJAUAN PUSTAKA

2.1 Data Mining

2.1.1 Pengertian Data Mining

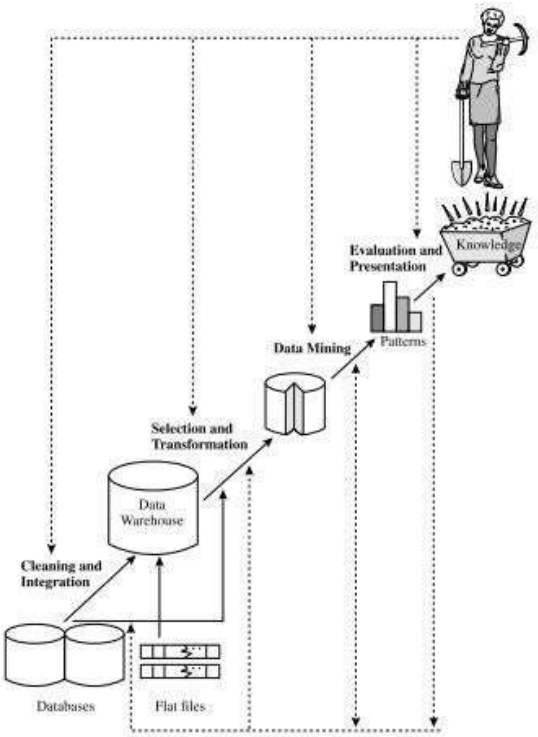
Secara sederhana, data mining dapat diartikan sebagai proses penggalian atau penemuan informasi dalam suatu kumpulan data. Informasi yang diperoleh dapat digunakan dalam berbagai bidang seperti manajemen bisnis, pendidikan, kesehatan, dll. Menurut Tacbir, data mining adalah suatu proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin untuk mengekstraksi informasi yang berguna dan informasi yang relevan untuk diekstraksi. dan mengidentifikasi database besar. Istilah data mining mengacu pada suatu disiplin ilmu yang tujuan utamanya adalah menemukan, memeriksa, atau mengekstrak informasi dari data yang ada. Proses penambangan data melibatkan integrasi teknik dari berbagai disiplin ilmu, seperti teknologi basis data dan gudang data, statistik, pembelajaran mesin, komputasi kinerja tinggi, pengenalan pola, jaringan saraf, visualisasi data, dll. Penambangan data menggunakan pendekatan berbasis penemuan yang menggunakan pencocokan pola dan algoritma lain untuk menentukan hubungan kunci dalam data yang diperiksa. Seperti namanya, data mining berarti mencari informasi bisnis yang berharga dalam database yang sangat besar. Ketika database dengan kualitas dan ukuran yang memadai tersedia, teknologi data mining memiliki karakteristik sebagai berikut[8]:

a. Otomatiskan perkiraan tren untuk fungsi bisnis. Penambangan data mengotomatiskan pengambilan informasi dari database besar.
b. Otomatiskan deteksi pola yang sebelumnya tidak diketahui. Alat penambangan data "mencari" database dan kemudian mengidentifikasi pola yang sebelumnya tersembunyi dalam

satu langkah. Contoh untuk menemukan pola ini adalah menganalisis data penjualan ritel untuk mengidentifikasi produk yang tampaknya tidak berkaitan yang sering dibeli pelanggan pada saat yang bersamaan.

2.1.2 Tahapan dalam Data Mining

Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahapproses yang diilustrasikan pada Gambar 3 Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan *knowledge base*[5].



Gambar 2.1 Tahap-tahap *Data Mining*

Tahap-tahap *data mining* adalah sebagai berikut:

a. Pembersihan data (*data cleaning*)

   Pembersihan data merupakan proses menghilangkan *noise* dan data yang tidakkonsisten atau data tidak relevan.

b. Integrasi data (*data integration*)

   Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru.

c. Seleksi data (*data selection*)

Data yang ada pada *database* sering kali tidak semuanya dipakai, oleh karena ituhanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.

d. Transformasi data (*data transformation*)

Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *datamining*.

e. Proses *mining*

Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data.

f. Evaluasi pola (*pattern evaluation*)

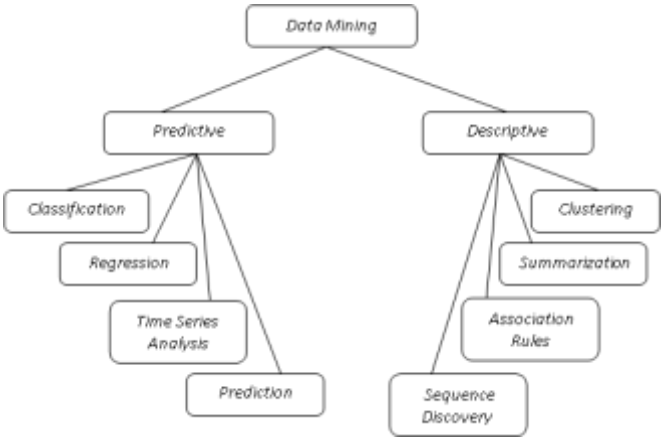
Untuk mengidentifikasi pola-pola menarik ke dalam *knowledge based* yangditemukan.

g. Presentasi pengetahuan (*knowledge presentation*)

Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

2.1.3 Teknik – Teknik Data Mining

Data mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data berupa pengetahuan yang selama ini tidak diketahui secara manual. Perlu diingat bahwa kata *mining* sendiri berarti usaha untuk mendapatkan sedikit data berharga dari sejumlah besar data dasar. Karena itu *data mining* sebenarnya memiliki akar yang panjang dari bidang ilmu seperti kecerdasan buatan (*artificial intelligent*), *machine learning*, statistik dan basis data[8]. Teknik *data mining* biasanya terbagi dalam dua kategori, prediksi dan deskripsi. Teknik prediksi menggunakan data historis untuk menyimpulkan sesuatu tentang kejadian di masa depan. Sedangkan teknik deskripsi bertujuan untuk menemukan pola dalam data yang menyediakan beberapa informasi tentang hubungan interval yang tersembunyi[8].



Gambar 2. 2 Teknik – Teknik Data Mining

Terdapat beberapa teknik yang digunakan dalam *data mining*, yaitu[8]:

a. Classification

Klasifikasi adalah teknik yang paling umum diterapkan pada *data mining*. Pendekatan ini sering menggunakan keputusan pohon (*decision tree*) atau *neural network* berbasis algoritma klasifikasi. Proses klasifikasi data melibatkan *learning* dan klasifikasi. Dalam belajar (*learning*) data pelatihan (*training*) dianalisis dengan alogaritma klasifikasi.

b. Clustering

Clustering bisa dikatakan sebagai identifikasi kelas objek yang memiliki kemiripan. Dengan menggunakan teknik clustering kita bisa lebih lanjut mengidentifikasi kepadatan dan jarak daerah dalam objek ruang dan dapat menemukan secara keseluruhan pola distribusi dan korelasi antara atribut. Pendekatan klasifikasi secara efektif juga dapat digunakan untuk membedakan kelompok atau kelas objek.

c. Predication

Teknik regresi dapat disesuaikan untuk predikasi. Analisis regresi dapat digunakan untuk model hubungan antara satu atau lebih independent variables dan dependent variables. Dalam data mining independent variabel adalah atribut-atribut yang sudah dikenal dan respon variabel apa yang kita inginkan untuk diprediksi.

d. Association rule

Digunakan untuk mengenali kelakuan dari kejadian-kejadian khusus atau proses dimana link asosiasi muncul pada setiap kejadian. Contoh dari aturan assosiatif dari analisa pembelian di suatu pasar swalayan adalah bisa diketahui berapa besar kemungkinan seorang pelanggan membeli roti bersamaan dengan susu. Dengan pengetahuan tersebut pemilik pasar swalayan dapat mengatur penempatan barangnya atau merancang kampanye pemasaran dengan memakai kupon diskon untuk kombinasi barang tertentu.

e. *Neural network*

Jaringan saraf adalah seperangkat unit penghubung input dan output dimana setiap koneksinya memiliki bobot. Selama fase learning, jaringan belajar dengan menyesuaikan bobot sehingga dapat memprediksi kelas yang benar label dari setiap input.

f. *Decision trees*

Decision trees atau pohon keputusan adalah struktur tree-shaped yang mewakili set keputusan. Keputusan ini menghasilkan aturan untuk klasifikasi sebuah kumpulan data. Metode pohon keputusan diantaranya yaitu Classification and regression trees (CART) dan Chi Square Automatic Interaction Detection (CHAID).

g. *Nearest Neighbor Method*

Teknik yang mengklasifikasikan setiap record dalam sebuah kumpulan data berdasarkan sebuah kombinasi suatu kelas k record yang sama dalam sebuah kumpulan data historis (dimana k lebih besar atau sama dengan 1). Terkadang disebut juga dengan teknik K-Nearest Neighbor.

2.2 Analisis Cluster

Analisis cluster dikenal juga sebagai teknik data mining dimana teknik ini bertugas dalam mengelompokkan suatu peristiwa dalam bentuk kelompok - kelompok kecil dimana setiap kelompok memiliki kemiripan objek satu sama lain. Jarak Euclidean disebut juga sebagai ukuran jarak yang digunakan pada analisis cluster [9].

$$x = (x_1, x_2, \dots, x_p) \text{ dan } y = (y_1, y_2, \dots, y_p) \text{ maka} \qquad (1)$$

$$d\,(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

$$d\,(x, y) = \sqrt{(x - y)'\,(x - y)'}$$

$$d\,(x, y) = \sqrt{\sum_{i=0}^{p} (x_i - y_i)^2}$$

Dimana: $d\,(x, y)$ = jarak antara objek x ke objek y

xi = data x pada observasi ke-i

yi = titik pusat ke y observasi ke-i

p = banyaknya variabel yang diamati.

2.3 Algoritma K-means

Pengertian dari *K-Means Clustering* adalah, K dimaksudkan sebagai konstanta jumlah *Clustering* yang diinginkan, *Means* dalam hal ini berarti nilai suatu rata-rata dari suatu grup data yang dalam hal ini didefinisikan sebagai *Clustering*, sehingga *K-Means Clustering* adalah suatu metode penganalisaan data atau metode Data Mining yang melakukan proses pemodelan tanpa *supervisi (unsupervised)* dan merupakan salah satu metode yang melakukan pengelompokan data dengan sistem partisi [9].

K-means sendiri adalah metode *cluster* data non-hierarki yang bertujuan mempartisi data dalam bentuk *cluster*. Data yang telah diperoleh memiliki karakteristik seimbang yang selanjutnya akan dikelompokan pada satu *cluster* yang sama.

Adapun algoritma yang digunakan dalam proses pengelompokkan data adalah sebagai berikut [9]:

a. Mentukan banyaknya cluster (nilai k).
b. Menentukan centroid (titik pusat) secara acak dibagi menjadi beberapa kelompok.

c. Menghitung pusat kelompok dari data yang ada. Lokasi kelompok diambil dari rata – rata semua nilai dari setiap fiturnya. Nilai n menyatakan banyaknya jumlah data yang didapat dalam sebuah kelompok, i menyatakan fitur ke-i dalam sebuah kelompok, dan p menyatakan dimensi data, maka persamaan untuk menghitung centroid fitur ke-i adalah sebagai berikut:

$$d\,(x, y) = \|x - y\|^2 \sqrt{\sum_{i=0}^{n} (x_1 - y_1)^2} \qquad (2)$$

Jarak adalah akar dari jumlah kuadrat dari pengurangan keanggotaan data ke-k dengan cluster ke- i dan nilai centroid cluster ke-i.

Dimana:

$d(x, y)$= jarak data ke x ke pusat cluser y

xi= data x pada observasi ke-i

yi= titik pusat ke y observasi ke-i

n= banyak observasi.

d. Data kemudian dialokasikan ke centroid/rata-rata terdekat. Diantaranya Eucliden pengukuran jarak sebagai berikut.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad (3)$$

Dimana:

d= jarak data ke x ke pusat cluser$y$

$(x_1 - y_1)$= kordinat titik pertama

$(x_2 - y_2)$= kordinat titik kedua

Selanjutnya data yang dihasilkan akan dialokasi kembali dengan menerapkan metode kmeans yang berdasarkan pada perbandingan jarak antara data dengan titik pusat pada setiap kelompok yang diperoleh.

$$a_{ii} = \{ \begin{matrix} 1, \\ 0 \end{matrix} \quad d = min\{D(x_{i,}\ C_1)\}\ lainnya \tag{4}$$

Dimana:

a_{ii}= variable atau elemen dalam suatu konteks

d= fungsi atau metrik jarak

D= metrik jarak yang digunakan untuk mengukur jarak antara dua titik

x_1= koordinat atau nilai dari dimensi pertama

C_1= pusat atau centroid dari klister pertama

Fungsi tujuan yang digunakan dalam metode k-means ditentukan berdasarkan jarak dan nilai keanggotaan data dalam kelompok.

e. Objek yang diperoleh kemudian dihitung kembali dengan menggunakan centroid baru. Proses clustering dikatakan selesai jika tidak terjadi perubahan pada anggota cluster yang ada.

2.4 Meode Elbow

Metode *ElbowMetode elbow*(siku) adalah sebuah pendekatan yang digunakan dalam analisis klastering untuk menentukan jumlah optimal dari klaster dalam data. Jika terdapat sudut atau penurunan yang paling signifikan antara nilai klaster pertama dan klaster kedua dalam grafik, maka nilai klaster tersebut dapat dianggap yang terbaik[10]

2.5 Repidminer

RapidMiner merupakan perangakat lunak yang bersifat terbuka (open source). RapidMiner adalah sebuah solusi untuk melakukan analisis terhadap data mining, text mining dan analisis prediksi. RapidMiner menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik. RapidMiner memiliki kurang lebih 500 operator data mining, termasuk operator untuk input, output, data preprocessing dan visualisasi. RapidMiner merupakan software yang berdiri sendiri untuk analisis data dan sebagai mesin data mining yang dapat diintegrasikan pada produknya sendiri. RapidMiner ditulis dengan munggunakan bahasa java sehingga dapat bekerja di semua sistem operasi[11]

2.6 Penelitian Terdahulu

Dalam melakukan penelitian ini terdapat penelitian yang telah dilakukan sebelumnya yang menjadi bahan referensi bagi peneliti. Berikut adalah penelitian terdahulu yang digunakan sebagai referensi, penelitian terdahulu dapat dilihat pada tabel 2.1.

Tabel 2. 1 Penelitian Terdahulu.

No	Peneliti	Judul	Metode	Kesimpulan
1.	Timothy Solang , Adi Nugroho	Analisis Kesehatan Mental Mahasiswa Universitas Kristen Satya Wacana Menggunakan Metode Clustering Algoritma K-Means	Algoritma K-Means	Berdasarkan hasil yang didapat cluster 1 berisi 11 data 34.38% untuk cluster 2 berisi 12 data 37.50% dan cluster 3 berisi 9 data 28.12%. Yang terdiri dari angkatan 2018, 2019, dan 2020 dapat tersimpulkan bahwa angkatan memiliki pengaruh dalam hal tertekan secara mental dalam lingkup Mahasiswa. Berdasarkan perhitungan k-means dipenelitian ini kesehatan mental dapat dikelompokan dengan mudah.
2.	Muhammad Rafi Nahjan , Nono Heryana, Apriade Voutama	Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj	Algoritma K-Means	Dengan metode Clustering K-Means dipakai untuk mengelompokan jenis transaksi yang sangat laris, laris dan kurang laris. Yang membuat kita tahu apa saja provider atau paket data yang digemari pada masyarakat sekitar Oj Cell, pen

		Cell		elitian ini juga memakai aplikasi RapidMiner agar mempermudah dan membuat hasil lebih akurat
3.	Tanty, Budi Serasi Ginting, Magdalena Simanjuntak	Pengelompokan Penyakit Pada Pasien Berdasarkan Usia Dengan Metode K-Means Clustering (Studi Kasus : Puskesmas Bahorok)	Algoritma K-Means	Peneliti melakukan proses menerapak k-means dalam pengelompokan penyakit pada pasien berdasarkan usia. Hasil dari penelitian ini dapat dikelompokan menjadi tiga cluster. Claster 1 pasien merupakan anak – anak, Cluster 2 pasien merupakan orang dewasa, Cluster 3 pasien merupaka orang lansia.

Penelitian yang dilakukan oleh Muhammad Rafi Nahjan, Nono Heryana, dan Apriade Voutama pada tahun 2023 dengan judul Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell. Penelitian ini mengelompokan jenis transaksi yang berdasarkan kategori sangat laris, laris dan kurang laris, mendapatkan 3 cluster yaitu cluster 2 sangat diminati, cluster 1 kurang diminati, cluster 0 cukup.

Berdasarkan Penelitian yang dilakukan oleh Tanty, Budi Serasi Ginting, dan Magdalena Simanjuntak pada tahun 2021, dengan judul Pengelompokan Penyakit Pada Pasien Berdasarkan Usia Dengan Metode K-Means Clustering (Studi Kasus : Puskesmas Bahorok). Penelitian ini mengelompokan penyakit pada pasien berdasarkan usia menggunakan metode K-Means Clustering dengan hasil cluster 1 pasien merupakan anak – anak, cluster 2 pasien merupakan orang dewasa, cluster 3 pasien merupakan Lansia.

Melalui penelitian yang telah dilakukan sebelumnya, terbukti bahwa algoritma K-Means mampu mengelompokkan data dengan efektif, seperti yang tergambar dalam analisis kesehatan mental Mahasiswa, penjualan toko, dan pengelompokan penyakit berdasarkan usia. Sementara penelitian terdahulu telah menggambarkan keberhasilan algoritma K-Means dalam membentuk kluster dari data yang beragam.

Penelitian saya terfokus pada implementasi kesehatan mental pada Mahasiswa dengan mengintegrasikan faktor-faktor seperti akademik, kesehatan fisik, dan dukungan sosial, penelitian ini berupaya untuk melihat hubungan yang lebih dalam antara variabel-variabel ini dengan kesehatan mental Mahasiswa. Diharapkan dengan metode K-Means akan terungkap pola-pola yang lebih spesifik mengenai bagaimana faktor-faktor tersebut berkontribusi pada kesehatan mental Mahasiswa.